New Pages 9 and 10:

distance between z and w_i (or between a and w_i) is approximated to be the distance between the midpoints of the intervals in which they lie. Also, the density function $f_X(a)$ is approximated to be the average of the density function in the interval in which the attribute "a" lies.

With this in mind,

$Pr'(X \in I_p) = (1/n) \Sigma$ (over s=1 to m) of $\{N(I_s) \times [(f_Y(m(I_s)-m(I_p))Pr(X \in I_p))] / [\Sigma$(over t=1 to m) of $(f_Y(m(I_s)-m(I_t))Pr(X \in I_t))]$, where

A1

I(x) is the interval in which "x" lies, $m(I_p)$ is the midpoint of the interval I_p, and $f(I_p)$ is the average value of the density function over the interval I_p, p=1,...m.

Using the preferred method of partitioning into intervals, the step at block 46 can be undertaken in $O(m^2)$ time. It is noted that a naive implementation of the last of the above equations will lead to a processing time of $O(m^3)$; however, because the denominator is independent of I_p, the results of that computation are reused to achieve $O(m^2)$ time. In the presently preferred embodiment, the number "m" of intervals is selected such that there are an average of 100 data points in each interval, with "m" being bound $10 \leq m \leq 100$.

It is next determined at decision diamond 48 whether the stopping criterion for the iterative process disclosed above has been met. In one preferred embodiment, the iteration is stopped when the reconstructed distribution is statistically the same as the original distribution as indicated by a X^2 goodness of fit test. However, since the true original distribution is not known, the observed randomized distribution (of the

perturbed data) is compared with the result of the current estimation for the reconstructed distribution, and when the two are statistically the same, the stopping criterion has been met, on the intuition that if these two are close, the current estimation for the reconstructed distribution is also close to the original distribution.

When the test at decision diamond 48 is negative, the integration cycle counter "j" is incremented at block 50, and the process loops back to block 46. Otherwise, the process ends at block 52 by returning the reconstructed distribution.

A1

Now referring to Figure 5, the logic for constructing a decision tree classifier using the reconstructed distribution is seen. Commencing at block 54, for each attribute in the set "S" of data points, a DO loop is entered. Moving to block 56, split points for partitioning the data set "S" pursuant to growing the data tree are evaluated. Preferably, the split points tested are those between intervals, with each candidate split point being tested using the so-called "gini" index set forth in Classification and Regression Trees, Breiman et al., Wadsworth, Belmont, 1984. To summarize, for a data set S containing "n" classes (which can be predefined by the user, if desired) the "gini" index is given by $1-\Sigma p_j^2$, where p_j is the relative frequency of class "j" in the data set "S". For a split dividing "S" into subsets S1 and S2, the index of the split is given by:

index = $n_1/n(\text{gini}(S1)) + n_2/n(\text{gini}(S2))$, where n_1 = number of classes in S1 and n_2 = number of classes in S2.

The data points are associated with the intervals by sorting the values, and assigning the $N(I_1)$ lowest values to the first interval, the next highest values to the next interval, and so on.